Randomized Sampling: An Approach to Extraction of Metadata Records

Olajumoke Azogu
University of North Texas
olajumoke.azogu@unt.edu

Jiangping Chen
University of North Texas
jiangping.chen@unt.edu

Abstract

Random samples are desired by researchers to produce results that are generalizable. We present an approach to extract random samples of metadata records from Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) repositories and others such as Z39.50 collections. Based on our approach, we developed PHP scripts to randomly extract 2000 metadata records from two digital repositories. Our approach can be used by others that need to extract representative samples from digital collections.

Keywords: algorithm, metadata, harvesting, extraction, oai-pmh, Z39.50

Introduction

Random sampling remains one of the preferred sampling methods for scientific research because it can generate representative samples of populations of interest, therefore guarantees the generalizability of results. With ever increasing digital objects and growing interests in investigating better digital content management, extracting representative samples through random sampling is important and efficient for working with large collections.

We present our approach to randomly extract any number of metadata records from digital collections. This approach was used to extract random samples for the MRT Project[1] from the Catalog of the University of North Texas Libraries (UNTL) and the Portal to Texas History (PTH) digital library. These two digital collections use different metadata standards.

Metadata Harvesting and Extraction Approaches

Our approach stemmed from existing methods for metadata harvesting; the process by which digital collections' metadata records can be accessed and extracted through established interchange standards among collections that enable sharing (Arms, Dushay, Fulker & Lagoze, 2003). We explored extraction across the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Z39.50 protocol.

OAI-PMH-based Harvesting/Extraction

The OAI-PMH developed for interoperability between digital collections (http://www.openarchives.org/pmh/), has been broadly acknowledged as a means for metadata harvesting (Van de Sompel, Nelson, Lagoze & Warner, 2004). It supports the identification and extraction of metadata records, and dissemination of records in various metadata formats. However, to ensure

[1] The title of this project is "Enabling Multilingual Information Access to Digital Collections: An Investigation of Metadata Records Translation," a research project sponsored by the Institute of Museum and Library Services (IMLS) and the University of North Texas (UNT) that aims to evaluate the extent to which current machine translation technologies generate adequate translation for metadata records, and to identify the most effective metadata records translation strategies for digital collections.

Acknowledgements: Institute of Museum and Library Services (IMLS) Grant LG-06-10-0162-10. We thank UNT Libraries for allowing us to use their digital collections. We also thank Mark Phillips and Jason Thomale for their help on this work.
Azogu, O., & Chen, J. (2013). Randomized sampling: An approach to extraction of metadata records. *iConference 2013 Proceedings* (pp. 804-807). doi:10.9776/13394

interoperability, unqualified Dublin Core is required and outputs from databases must be in XML (Warner, 2001).

Several approaches and tools exist that use OAI-PMH to extract metadata from a digital collection. Typical examples like MarcEdit (http://people.oregonstate.edu/~reeset/marcedit/html/index.php) and ZMARCO (http://zmarco.sourceforge.net/) either harvests all records, specified set of records, or an individual record in a collection (http://www.openarchives.org/pmh/tools/tools.php). Additional solutions for metadata extraction include approaches where all records from a digital collection are harvested and then systematic sampling applied to draw representative samples. However, these were found to be ineffective for the purposes of truly randomized metadata records extraction or inefficient when dealing with huge digital collections that contain millions of records, when only relatively few, representative number of records were needed.

Z39.50-based Harvesting/Extraction

Z39.50 was developed for the search and retrieval of information in databases (ANSI/NISO, 2003). A rapid assessment of Z39.50 tools (http://www.loc.gov/z3950/agency/ resources/software.html) reinforced the need for a flexible, adaptable application similar to that to be developed for the OAI-PMH-compliant catalog. Not unexpectedly, available Z39.50 applications, in a similar way to OAI-PMH tools, did not fulfill the needs of the MRT project.

The MRT project needed a solution that could effectively and efficiently extract specified numbers of records as a representative sample of a digital collection. In addition, it required the isolation of six (6) elements per metadata record. Existing OAI-PMH or Z39.50 tools could not achieve this. Our solution therefore contributes an original approach to the harvesting of metadata records, which provides users with control and flexibility over the number of records and metadata elements needed from digital collections. In this approach, we randomize extraction to ensure that representative metadata records are obtained irrespective of the number of records to be extracted from a collection.

Randomized Extraction of Sample Metadata Records

For the purposes of the MRT project, 2000 metadata records were to be extracted from the PTH and UNTL Catalogs. The extraction processed metadata records with two different standards: DC and MARC. As a result, six elements (Title, Publisher, Description, Subject & Keywords, Coverage and Creator) per metadata record were extracted.

The Approach

Our approach follows these steps, starting with the base URL of the collection;

(1) Obtain collections' metadata records identifier header and ID number range;
(2) Generate randomized IDs that fall within collection's range. Repeats if ID has been generated, else continues;
(3) Extract record with generated random ID;
(4) For the MRT project's unique need, six metadata elements of interest are isolated and saved;
(5) Repeat steps (2), (3) and (4) until desired numbers of records are extracted.

PHP, our scripting language was chosen for its functionalities to interface with both OAI-PMH and Z39.50 protocol servers. The above algorithm is illustrated in Figure 1.

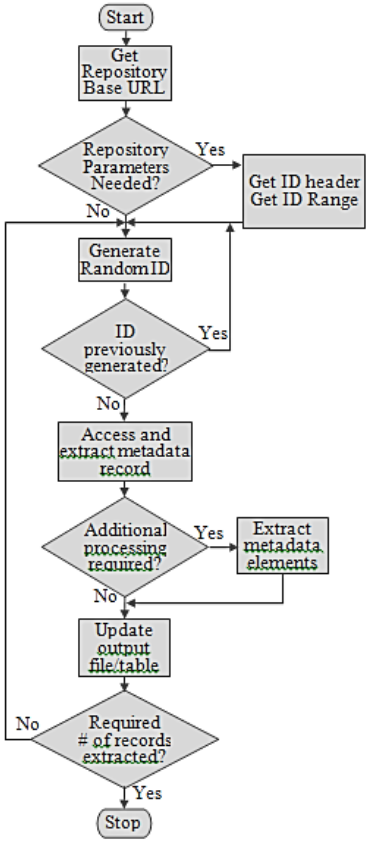

Figure 1. Randomized metadata records extraction algorithm

Extraction from the Portal to Texas History (PTH) Digital Collection

The OAI-PMH-compliant PTH collection, contained about 133,000 metadata records (http://texashistory.unt.edu/, January, 2011). Using OAI-PMH verbs, and PHP, we executed our algorithm to extract 1000 random metadata record samples.

Two PHP scripts were written for this process. While the repository's URL is needed and serves as entry point for the extraction process, the first script uses OAI-PMH verb ListIdentifiers to retrieve the repository's ID header, and lower / upper boundaries of the metadata records IDs. The information is passed on to the randomized extraction process as parameters. The second script uses the same parameters and the repository's URL to extract the number of records needed. It generates a random unique ID with the OAI-PMH verb GetRecord to access the corresponding metadata record. The extracted record is validated, six metadata elements obtained, multiple metadata element occurrences concatenated, and then written into database tables. Our script uses PHP's mt_rand() function (http://php.net/manual/en/function.mt-rand.php) to generate random numbers and validates metadata records by checking for non-empty records and for those containing at least four of the six elements of interest.

Extraction from the UNT Libraries Catalog

We adapt our algorithm to extract 1000 records from the 1.9 million records (January, 2011) Z39.50 UNTL Catalog. A script, using PHP and its YAZ extension (YAZPHP), obtains the repository URL, and runs similar to the OAI-PMH extraction except that YAZPHP functions (yaz_connect, yaz_search and yaz_record) are used for extracting records. In contrast also, upper and lower record ID limits were manually obtained and hard coded into the script.

The PHP scripts can be found at http://txcdk-v10.unt.edu/MRT/Scripts/welcome.html.

Experiments and Results

To affirm the validity and randomness of the extraction process, two experimental runs were carried out, and a comparative analysis of the results carried out. Runs were conducted over a Linux server via a command line interface, and outputs stored to MySQL tables and plain text files. The two sample batches of 1000 each, more than statistically significant sample sizes, were determined not to be significantly different, making us conclude that our random samples are a sufficient representative of the entire catalog metadata records.

Test runs on three OAI-PMH repositories and three Z39.50 repositories were successful. However, several OAI-PMH (http://gita.grainger.uiuc.edu/registry/ListAllRepos.asp) and Z39.50 repositories (http://www.loc.gov/z3950/) were inaccessible either because of broken links, or restrictions that require prior, written approval for access. Access is therefore important for successful metadata records extraction.

Discussion and Conclusion

We developed a generalizable solution for randomized metadata records extraction that is freely accessible. Other advantages include flexibility and control over the number of metadata records and elements extracted, and extensibility of the script to suit unique uses. While our scripts can be run “as-is”, repository-specific parameters would necessitate changes. However, our solution makes these necessary changes easy.

Despite our accomplishment, work still needs to be done to extend the generalizability of our solution. The command-line interface could be replaced with a user-friendly graphical user interface (GUI) which would negate the need for server-end scripting. Our work stemmed from the need for flexibility in extracting metadata records. Even greater flexibility may be achieved by merging our solution with other harvesting solutions, for such needs as; extracting random records over time periods for quality assurance, or, domain specific extractions which might be useful to research interests. Possibilities are numerous, and we conclude that adaptable metadata records extraction is unfinished.

References

ANSI/NISO (2003). Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Retrieved from http://www.loc.gov/z3950/agency/Z39-50-2003.pdf

Arms, W. Y., Dushay, N., Fulker, D. W. & Lagoze, C. (2003). A case study in metadata harvesting: The NSDL. *Library Hi Tech*, *21* (2), 228-237. doi:10.1108/07378830310479866

Van de Sompel, H., Nelson, M.L., Lagoze, C. & Warner, S. (2004). Resource Harvesting within the OAI-PMH Framework. *D-Lib Magazine, 10* (12). Retrieved from http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html

Warner, S. (2001). Exposing and harvesting metadata using the OAI metadata harvesting protocol: A tutorial. *High Energy Physics Libraries Webzine, 4*. Retrieved from http://library.cern.ch/HEPLW/4/papers/3/